

AD-A267 089



DIRECTORATE OF
HEALTH CARE STUDIES
AND CLINICAL INVESTIGATION

2
HJ

ASSESSMENT OF TWO DATA COLLECTION
APPROACHES FOR FORT BRAGG CHILD/ADOLESCENT
MENTAL HEALTH DEMONSTRATION PROJECT
USING POWER ANALYSIS

CR 93-002
PART I - FINAL REPORT

JULY 1993

UNITED STATES ARMY
MEDICAL DEPARTMENT CENTER AND SCHOOL
FORT SAM HOUSTON, TEXAS 78234-6100

DTIC
ELECTE
JUL 23 1993
S E D



STRIPED STATEMENT
Approved for public release
Distribution Unlimited

93-16603

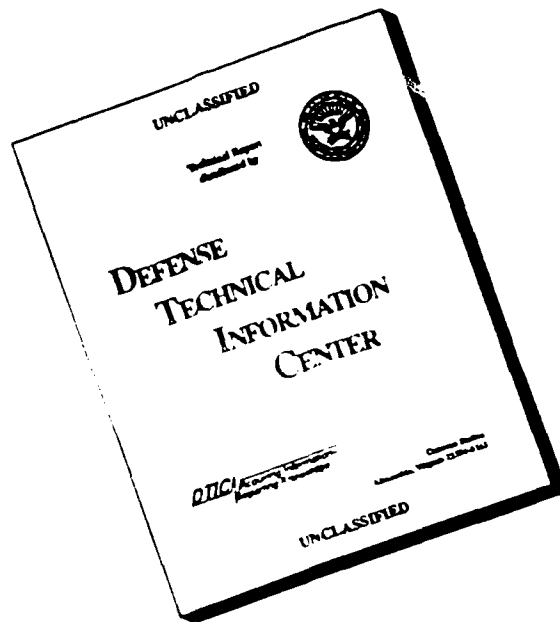


98

032

3808

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST
QUALITY AVAILABLE. THE COPY
FURNISHED TO DTIC CONTAINED
A SIGNIFICANT NUMBER OF
PAGES WHICH DO NOT
REPRODUCE LEGIBLY.

NOTICE

The findings in this report are
not to be construed as an official
Department of Defense position
unless so designated by other
authorized documents.

* * * * *

Regular users of services of the Defense Technical Information Center
(per DoD Instruction 5200.21) may purchase copies directly from the
following:

Defense Technical Information Center (DTIC)
ATTN: DTIC-DDR
Cameron Station
Alexandria, VA 22304-6145

Telephones: DSN 284-7633, 4, or 5
Commercial (703) 274-7633, 4, or 5

All other requests for reports will be directed to the following:

U.S. Department of Commerce
National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161

Telephone: Commercial (703) 487-4600

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION

Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Distribution Unlimited;
Public Use Authorized.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

CR93-002 Part I - Final Report

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Dir. Health Care Studies
and Clinical Investigation6b. OFFICE SYMBOL
(If applicable)
HSHN-A7a. NAME OF MONITORING ORGANIZATION
DASG

6c. ADDRESS (City, State, and ZIP Code)

Bldg 2268
Fort Sam Houston, TX 78234-6000

7b. ADDRESS (City, State, and ZIP Code)

Pentagon
Washington, D.C. 203018a. NAME OF FUNDING/SPONSORING
ORGANIZATION
HQ HSC8b. OFFICE SYMBOL
(If applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (City, State, and ZIP Code)

HQ HSC
Fort Sam Houston, TX 78234-6100

10. SOURCE OF FUNDING NUMBERS

PROGRAM
ELEMENT NO.PROJECT
NO.TASK
NO.WORK UNIT
ACCESSION NO.11. TITLE (Include Security Classification) (U) Assessment of Two Data Collection Approaches for
Fort Bragg Child/Adolescent Mental Health Demonstration Project Using Power
Analysis

12. PERSONAL AUTHOR(S)

Dr. Barbara E. Wojcik, Catherine R. Stein, M.S., Dr. Scott A. Optenberg

13a. TYPE OF REPORT
Final

13b. TIME COVERED

FROM Apr 93 to May 93

14. DATE OF REPORT (Year, Month, Day)

1993 JUL 02

15. PAGE COUNT

38

16. SUPPLEMENTARY NOTATION

This is a report to the Assistant Secretary of Defense (Health Affairs).

17. COSATI CODES

FIELD

GROUP

SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Fort Bragg Evaluation Project,
Statistical Power Analysis

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report presents the statistical review regarding an extension of the Fort Bragg Evaluation Project by Vanderbilt University Center for Mental Health Policy. It contains an assessment of two data collection plans using power analysis. The Monte Carlo power analysis performed by Vanderbilt University is also evaluated.

Based on the current short-term data collection plan submitted by the State of North Carolina, the statistical power is computed to be 80.258%. This level of power is considered high and should be adequate to meet the published Fort Bragg Evaluation Project statement of work.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT

☐ UNCLASSIFIED/UNLIMITED☒ SAME AS RPT☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL

Dr. Scott A. Optenberg

22b. TELEPHONE (Include Area Code)

(210) 221-0278

22c. OFFICE SYMBOL

HSHN-A

**ASSESSMENT OF TWO DATA COLLECTION
APPROACHES FOR FORT BRAGG CHILD/ADOLESCENT
MENTAL HEALTH DEMONSTRATION PROJECT
USING POWER ANALYSIS**

**A REPORT TO
THE ASSISTANT SECRETARY OF DEFENSE
(HEALTH AFFAIRS)**

**Dr. Barbara E. Wojcik, GM-13
Supervisory Statistician**

**Catherine R. Stein, MS, GS-11
Statistician**

**Dr. Scott A. Optenberg, GM-15
Chief, Health Care Analysis Division**

**Directorate of
Health Care Studies
and
Clinical Investigation**

**CR 93-002
Part I - Final Report
July 1993**

Accession For	
NTIS	CRA&I <input checked="checked" type="checkbox"/>
DTIC	TAB <input type="checkbox"/>
Unannounced <input type="checkbox"/>	
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

DTIC QUALITY INSPECTED 1

**UNITED STATES ARMY
MEDICAL DEPARTMENT CENTER AND SCHOOL
FORT SAM HOUSTON, TEXAS 78234-6100**

TABLE OF CONTENTS

SECTION	PAGE
DISCLAIMER	1
REPORT DOCUMENTATION PAGE	ii
TABLE OF CONTENTS	iv
BACKGROUND	1
POWER ANALYSIS COMPARISON OF TWO DATA COLLECTION PLANS	1
Power Analysis Assumptions	1
Power Analysis of Short and Long-Term Plans	3
Computational Procedure for the Exact Power of the Short and Long-Term Plans	4
Additional Power Computations	6
Assessment of the Simulation Method	7
CONCLUSION	9
REFERENCES	10
DISTRIBUTION LIST	12
APPENDIX A: LETTER DATED FEBRUARY 15, 1993, FROM DR. LENORE BEHAR TO MR. LEO SLEIGHT . . .	A-1 TO A-6
APPENDIX B: "STATISTICAL POWER IN CHILD PSYCHO- THERAPY OUTCOME RESEARCH," PAPER BY C. LAMPMAN, J. DURLAK, AND A. WELLS (PRESENTED AT 1992 AMERICAN PSYCHOLOGICAL ASSOCIATION CONVENTION)	B-1 TO B-2
APPENDIX C: POWER ANALYSIS DISCUSSION AND DOCUMENTATION FROM MATERIAL SUBMITTED BY VANDERBILT UNIVERSITY, APRIL 30, 1993	C-1 TO C-7
APPENDIX D: EXTERNAL PEER REVIEW OF THE REPORT CR 93-002	D-1 TO D-2

BACKGROUND

In response to inquiries from Congressional representatives, the Acting Assistant Secretary of Defense (Health Affairs) requested that the Army document a Department of Defense (DoD) position regarding an extension of the Fort Bragg Mental Health Demonstration Project. It was requested that the Army establish a panel of Army/DoD experts (psychiatrists, psychologists, other clinicians, and clinical statisticians) to review the evaluation and other related data concerning the Demonstration Project in order to: (1) support a DoD position on the level of confidence necessary to confirm treatment results/conclusions, and (2) indicate the impact of an Army approved evaluation due date on that level of confidence.

This technical report presents an independent statistical analysis/review. No actual data from the Fort Bragg Child/Adolescent Mental Health Demonstration Project or the Fort Bragg Evaluation Project were made available. However, information contained in a letter (shown as Appendix A) written by Dr. Lenore Behar, Ph.D., Head of the Child and Family Services Branch, North Carolina Department of Human Resources, to Mr. Leo Sleight, Central Contracting Office, Department of the Army, Headquarters U.S. Army Health Services Command, Fort Sam Houston, Texas, dated February 15, 1993, was provided by Vanderbilt University. In the letter, Dr. Behar presented two data collection plans. These plans, one Short-Term and one Long-Term, differ in the number of cases collected at 'Wave 3'. The effectiveness of each plan was described by means of a power value of a statistical test for detecting differences in improvement in mental health outcomes between Demonstration and Comparison cases. In addition, a reprint of a paper submitted to the 1992 American Psychological Association Convention addressing power analysis in psychotherapy research was furnished. This paper is included as Appendix B.¹ Also submitted was documentation supporting the power values in Appendix A in materials attached to a letter dated April 30, 1993, written by Dr. Leonard Bickman, Ph.D., Director of the Center for Mental Health Policy, Institute for Public Policy Studies, Vanderbilt University, to LTC Thomas E. Leonard, Headquarters U.S. Army Health Services Command, Fort Sam Houston, Texas. Pertinent portions of this documentation are included as Appendix C.

POWER ANALYSIS COMPARISON OF TWO DATA COLLECTION PLANS

Power Analysis Assumptions.

In the statistical assumptions presented in Appendix A, the type of variable(s) used to measure 'improvement' between an average Demonstration case and an average Comparison case was

not defined. The paper shown in Appendix B was referenced instead, presenting the results of a meta-analysis for 12 categories of outcome measures, six each for behavioral and nonbehavioral treatments. It appears that the Fort Bragg Evaluation Project used the Appendix B paper to obtain the value of the effect size (ES) for Normed Rating Scales--Nonbehavioral Treatment outcome measures--as this value is included in Appendix A. In Appendix A (p. A-6), it is stated that the Short-Term Plan has 50% power and the Long-Term Plan of data collection would have 80% power. These levels of power were based on a simulation model submitted by Vanderbilt University (Appendix C).

The effect size (ES) index identified as d by Cohen (1988),² is the standardized difference between two population means. This equation is as follows:

$$d = \frac{m_A - m_B}{\sigma}$$

where d = ES index for t test of means,
 m_A, m_B = population means,
 and σ = standard deviation of either population
 (equal variance is assumed).

The effect size value (ES = 0.25) derived in Appendix B (p. B-2) and cited in Appendix A (p. A-5) should be used with caution for several reasons. First, this value was computed for a series of 12 sub-group samples. The Normed Rating Scale used to derive the power in Appendix A was based on a mean sample of only 33 cases. The authors of the Appendix B paper stated this problem of variability as follows (p. B-2): "The large discrepancies between sample sizes actually used and those necessary to attain an acceptable level of power in the studies shown in Table 1 make it difficult to assess how closely the obtained treatment effect sizes represent true population effects. This, in turn underscores the need for researchers to attend to power considerations when planning therapy outcome studies." When a meta-analysis is based on such a small size the probability of error is high. As a result, the mean effect size (ES = 0.25) used in Appendix A may or may not express score distances (in units of variability) for the actual variables measuring health outcome in the Fort Bragg Evaluation Project.

Secondly, there is always a risk that meta-analysis may have employed inappropriate assumptions with regard to the validity of pooling and generality. For instance, the meta-analysis may contain some bias as to how the outcome should be produced, excluding some relevant trials from analysis. In other instances, meta-analysis may use multiple results from the same study, and because the results are not independent they may

bias or invalidate the meta-analysis. In other cases, the independent studies may include different measuring techniques and definitions of variables, so the outcomes may not be comparable. In general, effect sizes in unique areas are likely to be small ($ES = 0.20$ or $ES = 0.30$), but only a pilot test would give an answer as to the probable magnitude of the ES index for the particular variable of interest in a particular situation.

The power and sample size tables (Cohen, 1988)³ for the above specified $ES = 0.25$ in Appendix A are designed to analyze the difference between means of two independent samples of the same size drawn from normal populations with equal variances (using the t test for means). If these assumptions cannot be made, which is often the case, the additional adjustments that follow are explicitly supported by Cohen (1988)⁴ and others. Computations should be performed to obtain the harmonic mean if samples of different sizes but equal variance are present, and the root mean square should be computed if two samples of the same size having unequal variances are present. If both sample sizes and variances differ, the values for power formulas from the tables cited in Appendix A may not be valid.

Since no actual data were available from the Fort Bragg Evaluation Project, this review will utilize the data used by Vanderbilt University for this analysis. Appendix A contains a comparison of the two data collection plans using power analysis. The Appendix A power analysis comparison presents the number of cases after attrition for both the Short-Term and Long-Term Plans (p. A-6). For the Short-Term Plan, 299 Demonstration cases and 150 Comparison cases were expected. The following power analysis is based on Cohen's formulas and uses the information supplied in Appendix A. This analysis is followed by a discussion of the simulation submitted by Vanderbilt University and included as Appendix C.

Power Analysis of Short and Long-Term Plans.

Under the assumption that the variances in the Demonstration and Comparison sites are equal, the harmonic mean (n) of the Demonstration sample size (n_D) and the Comparison sample size (n_C) is given by the formula (Cohen, 1988):⁵

$$n = \frac{2n_D n_C}{n_D + n_C} = \frac{2(299)(150)}{299 + 150} = \frac{89,700}{449} \approx 200.$$

The value for power of the t test of the Demonstration case mean (m_D) and the Comparison case mean (m_C) testing the null hypothesis that $m_D = m_C$ at $\alpha_1 = 0.05$ (one-tailed test) (Table 2.3.2 from Cohen, 1988)⁶ gives the following results:

for $n = 200$ and $ES = 0.20$, power = 0.64, and
 for $n = 200$ and $ES = 0.30$, power = 0.91.

The effect size, proposed in Appendix A and derived from a meta-analysis performed in Appendix B, is 0.25. A linear interpolation was performed to derive the power of the t test for $ES = 0.25$. This computation yielded a power of 0.78 for $ES = 0.25$, $\alpha_1 = 0.05$ and $n = 200$. This power of 0.78 (78%), as computed for the Short-Term Plan, is much higher than the 0.50 (50%) quoted in Appendix A. A full precision computation of the power for the Short and Long-Term Plans is presented in the next section of this report.

The Long-Term Plan projects 426 Demonstration cases and 361 Comparison cases. This harmonic mean, computed under the assumption that the variances are the same, is as follows (Cohen, 1988):⁷

$$n = \frac{2n_D n_C}{n_D + n_C} = \frac{2(426)(361)}{426 + 361} = \frac{307,572}{787} = 390.8 \approx 391.$$

Employing Table 2.3.2 in Cohen (1988),⁸ $n = 350$ yields power = 84% for $ES = 0.20$ and power = 99% for $ES = 0.30$. For $n = 400$, power = 88% for $ES = 0.20$ and power is greater than 99% for $ES = 0.30$. The linear approximation yields a power of 93.3% for $ES = 0.25$ (for $n = 391$).

Computational Procedure for the Exact Power of the Short and Long-Term Plans.

The linear interpolation to compute power, discussed on pages 3 and 4, was justified by its simplicity and by the relatively accurate values obtained. The full precision in computing the power for the Short and Long-Term Plans was based on the expression (Cohen, 1988):⁹

$$z_{1-\beta} = \frac{d(n-1)\sqrt{2n}}{2(n-1) + 1.21(Z_{1-\alpha_1} - 1.06)} - z_{1-\alpha_1}$$

where $z_{1-\beta}$ = the percentile of the standard normal distribution giving the power value
 $z_{1-\alpha_1}$ = the percentile of the standard normal distribution for α_1 significance level
 d = the effect size ES
 and n = the harmonic mean.

For the Short-Term Plan, the following information was available:

$$\begin{aligned} n &= 200 \\ \alpha_1 &= 0.05 \\ d &= 0.25 \\ z_{1-\alpha_1} &= 1.645. \end{aligned}$$

The $z_{1-\beta}$ percentile was computed under these assumptions from the above formula:

$$\begin{aligned} z_{1-\beta} &= \frac{(0.25)(200 - 1)\sqrt{2(200)}}{2(200 - 1) + 1.21(1.645 - 1.06)} - 1.645 \\ &= \frac{(0.25)(199)(20)}{398 + (1.21)(0.585)} - 1.645 = \frac{995}{398.708} - 1.645 \\ &= 2.496 - 1.645 = 0.851. \end{aligned}$$

The probability for this $z_{1-\beta}$ percentile was found from the Normal Curve Areas Table C (Daniel, 1988).¹⁰ This probability presents the power of the test and is equal to 80.258%. The Short-Term Plan gives a statistical power (computed with full precision) exceeding 80%.

A similar computation was performed for the Long-Term Plan under the following assumptions:

$$\begin{aligned} n &= 391 \\ \alpha_1 &= 0.05 \\ d &= 0.25 \\ z_{1-\alpha_1} &= 1.645. \end{aligned}$$

The $z_{1-\beta}$ percentile found from the same formula (Cohen, 1988)¹¹ was computed as follows:

$$\begin{aligned} z_{1-\beta} &= \frac{(0.25)(391 - 1)\sqrt{(2)(391)}}{2(391 - 1) + 1.21(1.645 - 1.06)} - 1.645 \\ &= \frac{(97.5)(27.964)}{780 + 0.70785} - 1.645 = \frac{2,726.516}{780.708} - 1.645 \\ &= 3.492 - 1.645 = 1.847. \end{aligned}$$

The power for this value of $z_{1-\beta}$ found from the Normal Curve Areas Table C (Daniel, 1988)¹² is equal to 96.78%.

Additional Power Computations.

The power analysis shown above projects that the number of cases in the Short-Term Plan is currently sufficient to draw statistically significant conclusions with high statistical power (80.258%). An additional reason for this conclusion is found by using the sample size tables provided by Cohen (1988)¹³ and deriving the sample size necessary to achieve full 80% power. Sample size tables provide data for two homogeneous normally distributed populations from which random samples of the same size were derived. The ES specified in Appendix A is 0.25. This ES level is not tabulated by Cohen (1988).¹⁴ Therefore, to find the sample size for an untabulated effect size, the following formula is used (Cohen, 1988):¹⁵

$$n = \frac{n_{.10}}{100d^2} + 1$$

where $n_{.10}$ is the sample size for desired power,
given α and $ES = 0.10$,
and d is the effect size.

In addition, if the sample sizes are not equal, one sample size is treated as if fixed, while the other is computed. When the choice is arbitrary, it is generally supported that n_c be fixed and n_D be computed. To find n_D , the following formula is used (Cohen, 1988):¹⁶

$$n_D = \frac{n_c n}{2n_c - n}$$

where n_c = fixed sample size (Comparison sites),
 n = value read from the Table 2.4.1 (Cohen,
1988)¹⁷ or computed from the previous equation,
and n_D = sample size for the Demonstration site.

With the objective to determine the Demonstration case sample size required to yield a power = 80% with $\alpha_1 = 0.05$ and $ES = 0.25$, and fixing the Comparison cases at $n = 150$ (the current level), the formula for computing n is:

$$n = \frac{n_{.10}}{100d^2} + 1 = \frac{1,237^*}{100(0.25)^2} + 1 = \frac{1,237}{6.25} + 1 \approx 198 + 1 = 199.$$

*Source: Table 2.4.1 (Cohen, 1988).¹⁸

Next, this value is put into the formula for n_D :

$$n_D = \frac{n_C n}{2n_C - n} = \frac{(150)(199)}{2(150) - 199} = \frac{29,850}{300 - 199}$$

$$= \frac{29,850}{101} = 295.54 \approx 296.$$

Consequently, 296 Demonstration site patients are needed to assure an 80% power for the test investigating the difference in mental health outcomes between Demonstration and Comparison patients (299 were projected in Appendix A).

The identical procedure was applied to the Long-Term Plan. Given that the Comparison sites consist of 361 cases, and assuming the same conditions ($\alpha_1 = 0.05$, $ES = 0.25$, power = 0.80), a sample size of 138 cases for the Demonstration site was obtained:

$$n = \frac{n_{.10}}{100d^2} + 1 = 199$$

$$n_d = \frac{n_C n}{2n_C - n} = \frac{(361)(199)}{2(361) - 199} = \frac{71,839}{722 - 199} = \frac{71,839}{523}$$

$$= 137.36 \approx 138.$$

As proposed, in Appendix A, the Long-Term Plan is projected to produce 426 Demonstration cases. Using Vanderbilt University's information taken from Appendix A, the above analysis computes only 138 cases are statistically necessary to achieve 80% power.

Assessment of the Simulation Method.

Vanderbilt University's use of the Monte Carlo simulation method to perform a power analysis in the present situation is an inappropriate application of this type of simulation. Using simulation to compute the power analysis without any information about the actual data is not an appropriate use of either simulation or power analysis. Concerning simulation, Miller and Starr (1969)¹⁹ state:

"...Simulation is not a substitute for knowledge [emphasis by authors]. This cannot be over-emphasized. Simulation is not a method, which, somehow, compensates for lack of knowledge."

In general, simulation should be treated as a technique of "last resort" (Naylor, 1971),²⁰ to be used only when analytical techniques are not available for obtaining solutions to a given model. Power analysis gives the correct probability of getting a significant result of Comparison and Demonstration site means only when the effect size is computed precisely (i.e., based on actual data from actual variables in the experiment under consideration).

The use of simulation requires complete information about the process or object. In order to simulate reasonably, the probability distributions of the variables of interest should be known. If these distributions are not known, it is impossible to simulate the process. This position is strongly emphasized by many authorities in operations research (Naylor; Ignizio and Gupta; Buffa; Smith; Banks and Carson; Gibra; and Miller and Starr).²¹ It is critical that estimates of parameters of the simulation model be derived on the basis of observations taken from the actual data. Naylor (1971)²² states:

"... There is very little to be gained by using an inadequate model to carry out simulation experiments on a computer because we would merely be simulating our own ignorance."

Since the Monte Carlo technique presented in Appendix C does not involve actual data, the results obtained from this method may be entirely misleading and not accurate. The simulation shown in Appendix C is based on assumptions regarding the effect size ($ES = 0.25$). This value, derived from meta-analysis (Appendix B, p. B-2), may not apply to real differences between the mean values of mental health outcomes for the Demonstration and Comparison sites. Another assumption (Appendix A, p. A-5), regarding the average child improvement by 0.3 SD, due to treatment and time, is only theoretical because it is not based on actual data.

As stated above, Monte Carlo simulation should only be utilized when direct data analysis cannot be performed (Gibra, 1973),²³ which is not the case with the Fort Bragg Evaluation Project. In addition, the real probability distributions of all the random variables under consideration must be given (Gibra, 1973),²⁴ a fact ignored in Appendix C. The Monte Carlo method gives only approximations to sampling distributions (Snedecor and Cochran, 1980).²⁵ To this extent, the technique itself is subject to sampling error.

Another observation about the Appendix C discussion was that the Monte Carlo method was performed only for one variable (CBCL); no other variables were used. The analysis might had different results if the other variables were considered. Finally, before any simulation model can be accepted it must be verified and validated to identify model biases and erroneous assumptions, if any. The authors of the modeling as reported in Appendix C included no such validation.

Without the use of actual data, the effect size value (derived from the meta-analysis cited in Appendix B) was used to calculate the power in this report. This effect size was recommended by the staff of the Fort Bragg Evaluation Project. Although not considered actual data, the effect size allowed for no additional bias to be created by the Monte Carlo method. The equations used to compute the power of the test of means in this report are supported by numerous authorities in power analysis (Cohen, 1988).²⁶

CONCLUSION

The power values for the directional tests computed in this study and the values given in the proposal in Appendix A are significantly different. Utilizing information available in Appendix A and a methodology well supported in the statistical literature, this study demonstrates that the Short-Term Plan would yield power exceeding 80% (80.258%) at full precision, instead of 50% as presented in Appendix A. Even using linear interpolation, a power of 78% was derived. This study demonstrates that it is unnecessary to extend the duration of the project based on power requirements; the Short-Term Plan should produce high power to demonstrate significance if the alternative hypothesis is true. The Demonstration sample size n_D needed to achieve 80% power for the Short-Term Plan ($\alpha = 0.05$, $n_C = 150$, $ES = 0.25$) equals 296 cases.

Secondly, because the standardized effect size is a computed variable, it can be modified. This modification can be achieved by any of several methods currently available to the Fort Bragg Evaluation Project staff without any project extension. Variance can be reduced, thereby allowing a decrease in sample size necessary to detect a particular level of effect size at a specified power by increasing quality control in data collection and preparation for analysis. For example, each outcome should be used in as sensitive a form as can be reliably measured (variable of interest should always be measured on a continuum, not dichotomized). Unnecessary dichotomization causes a loss of power in all analyses. Consequently, a much larger sample is necessary to achieve the same power.

Finally, as stated above, a more accurate estimate of the Fort Bragg Evaluation Project effect size is achieved when actual data is utilized and a full post hoc power analysis is conducted. The advisability of performing post hoc power analysis is strongly supported by Cohen (1988),²⁷ Rossi (1990),²⁸ Bailer (1992),²⁹ and numerous authorities on power analysis in the behavioral/medical sciences.

REFERENCES

1. Claudia Lampman, Joseph Durlak, and Anne Wells, "Statistical Power in Child Psychotherapy Outcome Research," Paper presented at the annual convention of the American Psychology Association, 1992.
2. Jacob Cohen, Statistical Power for the Behavioral Sciences (Hillsdale, NJ: Lawrence Erlbaum Associates, 1988), 20.
3. Ibid.
4. Ibid., 42.
5. Ibid., 42.
6. Ibid., 31.
7. Ibid., 42.
8. Ibid., 31.
9. Ibid., 544.
10. Wayne W. Daniel, Essentials of Business Statistics, 2nd Ed. (Boston, MA: Houghton Mifflin Co., 1988), A26-A27.
11. Cohen, 544.
12. Daniel, A26-A27.
13. Cohen, 54.
14. Ibid., 54.
15. Ibid., 53.
16. Ibid., 59.
17. Ibid., 54.
18. Cohen, 54.
19. David W. Miller and Martin K. Starr, Executive Decisions and Operations Research, 2nd Ed. (Englewood Cliffs, NJ: Prentice-Hall, 1969), 556.
20. Thomas H. Naylor, Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, 1971).
21. Thomas H. Naylor, Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, 1971); James P. Ignizio and Jatinder N. D. Gupta, Operations Research in Decision Making, with the collaboration of Gerald R. McNichols

(New York: Crane, Russak & Co., 1975); Elwood S. Buffa, Operations Management: Problems and Models, 3rd Ed. (New York: John Wiley & Sons, 1972); V. Kerry Smith, Monte Carlo Methods: Their Role for Econometrics (Lexington, MA: Lexington Books, D.C. Heath and Co., 1973); Jenny Banks and John S. Carson, II, Discrete-Event System Simulation (New York: Prentice-Hall, 1984); Isaac Gibra, Probability and Statistical Inference for Scientists and Engineers (Englewood Cliffs, NJ: Prentice-Hall, 1973); and David W. Miller and Martin K. Starr, Executive Decisions and Operations Research, 2nd Ed. (Englewood Cliffs, NJ: Prentice-Hall, 1969).

22. Naylor, 14.

23. Isaac N. Gibra, Probability and Statistical Inference for Scientists and Engineers (Englewood Cliffs, NJ: Prentice-Hall, 1973), 43.

24. Ibid.

25. George W. Snedecor and William G. Cochran, Statistical Methods, 7th Ed. (Ames, IA: Iowa State University Press, 1980), 9.

26. Cohen.

27. Ibid., 14.

28. Joseph S. Rossi, "Statistical Power of Psychological Research: What Have We Gained in 20 Years?," Journal of Consulting and Clinical Psychology 58 (1992): 646-656.

29. John C. Bailar III and Frederick Mosteller, Medical Uses of Statistics, 2nd Ed. (Boston, MA: NEJM Books, 1992), 47.

DISTRIBUTION LIST

Administrator, Defense Technical Information Center, ATTN:
DTIC-OOC (Selection), Bldg 5, Cameron Station, Alexandria,
VA 22304-6145 (2)

Director, Joint Medical Library, DASG-AAFJML, Offices of
the Surgeons General, Army/Air Force, Rm 670, 5109 Leesburg
Pike, Falls Church, VA 22041-3258 (1)

Director, The Army Library, ATTN: ANR-AL-RS (Army
Studies), Rm 1A518, The Pentagon, Washington, DC 20310 (1)

Defense Logistics Studies Information Exchange, U.S. Army
Logistics Management College, Fort Lee, VA 23801-8043 (1)

Commandant, Academy Health Science, ATTN: HSHA-Z,
Fort Sam Houston, TX 78234-6100 (1)

Stimson Library, Academy of Health Sciences, Bldg 2840,
Fort Sam Houston, TX 78234-6100 (1)

Medical Library, Brooke Army Medical Center, Reid Hall,
Bldg. 1001, Fort Sam Houston, TX 78234-6200 (1)

The Assistant Secretary of Defense (Health Affairs), The
Pentagon, Washington, DC 20301-1200 (3)

Office of the Assistant Secretary of Defense (HA), Health
Services Financing (HSF), Coordinated Care Policy, Rm 1B657,
The Pentagon, Washington, DC 20301-1200 (3)

HQ HSC (HSCL-M), ATTN: COL Beumler, Fort Sam Houston, TX
78234-6000 (3)

HQ HSC (HSAA-C), ATTN: Ms Emily Mathis, Fort Sam Houston, TX
78234-6000 (3)

APPENDIX A

LETTER DATED FEBRUARY 15, 1993, FROM
DR. LENORE BEHAR TO MR. LEO SLEIGHT



North Carolina Department of Human Resources
Division of Mental Health, Developmental Disabilities
and Substance Abuse Services

325 North Salisbury Street • Raleigh, North Carolina 27603 • Courier # 56-20-24

James B. Hunt, Jr., Governor
C. Robin Britt, Secretary

Michael S. Pedneau, Director
(919) 733-7011

February 15, 1993

Mr. Leo Sleight
Central Contracting Office
HSAA-C, Building 2015
Department of the Army
Headquarters, U.S. Army Health Service Command
Fort Sam Houston, Texas 78234-6000

Re: DADA10-89-C-0013, Fort Bragg Child/Adolescent Mental Health
Demonstration Project; Extension of Evaluation Component.

Dear Leo:

We have reviewed the status of the Evaluation Component of the Fort Bragg Child/Adolescent Mental Health Demonstration Project and find that, in keeping with the contract and with the Vanderbilt Statement of Work, the following reports will be submitted by September 30, 1993:

1. Implementation Study, Final Report.
2. Quality Study, Final Report.
3. Cost Study, Interim Report. As explained in Attachment 1, the data to be used for an Interim Report of the Cost Study to be submitted in September 1993 will be for FY92. As this report will be prepared during the last quarter of FY93, CHAMPUS data after September 1992 would be unstable given the time lag between the date of service and the appearance of those costs in the data. Another reason for using FY92 cost data is that Gateway cost data for FY93 would not be available in an analyzable form for a September 1993 report.

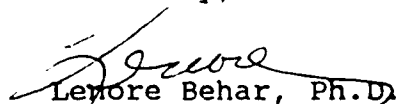
As explained in Attachment 2, it is not possible to complete the Outcome Study with an acceptable level of confidence by September 1993. If data were to be collected using the time frame proposed in the short term plan, the level of confidence, based on a sophisticated power analysis specific to this type of study, would be at the .50 level. As you know, this level of confidence is comparable to flipping a coin. We believe instead that the Outcome Study should be completed as originally designed to yield results at the .80 level of confidence. To achieve this goal,

Wave 1 data should be collected through June 30, 1993 at the Demonstration site and through December 31, 1993 at the Comparison sites; Wave 2 and Wave 3 data should be collected through June 30, 1994. Cost data specific to the clients in the study will need to be analyzed through the same time period in order to determine Cost Effectiveness. A final report of the Outcome Study and the companion Cost Effectiveness Study will be issued in September 1994. Costs of extending this portion of the Evaluation Component to completion are provided as Attachment 3.

It does not seem sensible to have invested in the Outcome Study thus far and terminate it short of having adequate information to reach a conclusion regarding the impact of the Demonstration Project on treatment outcomes. I will point out that no CHAMPUS evaluations in the past have addressed outcome, but rather have studied utilization and cost only. This absence of outcome data has been raised as a deficiency in the evaluation of the CPA Norfolk Demonstration (Burns, 1993, Attachment 4). I believe that the opportunity should not be prematurely abandoned to determine whether or not the methods of service delivery affect treatment outcomes. As we have discussed earlier, the delays which resulted from the failure of HSC to provide access to necessary data during the first two years of the project have seriously compromised the completion of the Outcome Study in a timely fashion. During those two years, I repeatedly emphasized the anticipated costliness of HSC's delays in providing access to data, so none of us should be surprised by the need to extend the Evaluation Component at this point.

As I noted in my letter dated December 16, 1992 (Attachment 5), I have discussed, with the various stakeholders, your plan to end the Evaluation Component before it is completed based on your belief that sufficient information exists to document the success of the project. I believe that those stakeholders maintain the same position now as they did in December; that is, that they wish to have unbiased and convincing evidence regarding this project and until such data are presented and accepted, we will need to continue the objective evaluation. I believe that this position is sound considering the issues from a scientific perspective. I trust you will endorse the merits of this position and support the completion of the Outcome Study and the Cost Effectiveness Study.

Sincerely,


Lenore Behar, Ph.D.

Head, Child and Family Services Branch

cc: Mr. James Newman

finalrep.hsc (13)

**PROPOSED CONTENTS OF COST STUDY OF
THE SECOND INTERIM REPORT (September 30, 1993)**

The Cost Study portion of the of the Fort Bragg Evaluation will assemble data from CHAMPUS records, Rumbaugh's management information system (MIS), Fort Campbell's MIS and Fort Stewart's medical records into an integrated utilization database. In addition, unit cost measures will be collected from each site, or estimated where not directly available. Using these data, Vanderbilt will produce an Interim Report to be submitted on September 30, 1993. In order to minimize bias due to start-up issues at the Demonstration, the report will be limited to the FY92 time period (October 1991 - September 1992). The lack of stability of CHAMPUS data for the period of time after September 1992 precludes inclusion of further data in the Interim Report. The Final Report, which will be submitted in September, 1994 will include subsequent cost data for all sites.

The Interim Report will provide a comparative analysis in tabular and/or graphic form, of the following:

<u>Service Category</u>	<u>Measures & Statistics</u>
Residential Services	\$ per day per eligible child
Hospital	\$ per day per child served
RTC	Admissions
Group & Therapeutic Home	Children served*
	Length of stay*
Non-Residential Services	\$ per day per eligible child
Day TX/In-Home	\$ per day per child served
Outpatient	Admissions
Medical Services	Length of episode*
(meds & med evals)	Number of episode*
Support Services	
(non-direct Services,	
e.g., Treatment Team	
activities & case man.	
phone calls)	

*Mean, median, maximum and minimum will be presented for these measures.

COMPARISON OF TWO DATA COLLECTION SCENARIOS USING POWER ANALYSIS

A power analysis was conducted to predict the consequences of ending data collection for the Evaluation before three waves of data could be collected on the targeted number of clients that is specified in the statement of work. Two plans have been discussed that differ in how long data would be collected. The objective of power analysis is to determine the number of cases for which data should be collected in order to determine the effectiveness of the Demonstration on children's mental health outcomes. The threat of collecting data from too few participants is that the statistical analysis may indicate that results were due to chance when, in fact, there was an undetected effect. Obviously if the final analysis misses the effect and tells us only that what we observed may be due to chance, the money and effort invested in the Evaluation and the Demonstration will have been wasted.

Power analysis is a specialized branch of psychological statistics that calculates how many subjects are needed to be assured that the results are not due to chance. It indicates "...the probability that statistical significance will be attained given there really is a treatment effect" (Lipsey, 1990, p. 20). To conduct power analyses, statisticians must make assumptions before calculating the proposed study's power. Those assumptions are discussed below.

Data Collection Assumptions

The power analyses presented here are based on two different data collection plans. They are as follows:

1. The short-term plan stops recruitment (Wave 1) at all sites on June 30, 1993, for a total of 1065 Wave 1 cases, and stops all data collection for Waves 2 and 3 on September 30, 1993. This plan, after correction for attrition, would include approximately 299 Demonstration cases and 150 Comparison cases with complete Wave 3 data.
2. The longer plan stops recruitment (Wave 1) at the Demonstration site on June 30, 1993, and at the Comparison sites December 30, 1993, for a total of 1125 Wave 1 cases. In this scenario, all Wave 2 and Wave 3 data collection would end June 30, 1994. This plan, after correction for attrition, would include approximately 426 Demonstration cases and 361 Comparison cases at Wave 3.

Clinical Assumptions

1. Many children will improve in both settings, but more children will improve in the Demonstration because there will be, on the average, a better fit between the child and his or her treatment.
2. Important improvement due to the treatment will continue to accrue for at least the first year following the start of treatment.

Statistical Assumptions

1. Statistical tests will be run at $p(\alpha) = 0.05$. This means we will follow the scientific norm of being 95% certain that observed differences are not due to chance.
2. The Evaluation is attempting to detect a difference of at least 0.25 standard deviations (SD) difference in improvement between the average child at the Demonstration site and the average child at the Comparison sites. The study will not be capable of effectively detecting effects smaller than .25 SD. This effect size was derived from a meta-analysis on child psychotherapy as the mean effect size found for nonbehavioral treatment using instruments similar to ones used in the Evaluation (Lampman, Durlak & Wells, 1992). A difference of 0.25 SD is the same as saying that if 50% of the patients in the Comparison get better while 63% of those in the Demonstration will get better.
3. All children, on the average, will improve by 0.3 SD due to treatment and time; Demonstration children will improve an additional 0.25 SD due to treatment conditions unique to the Demonstration.
4. The goal is to determine only if the children at the Demonstration site have better outcomes, in general, than the children at the Comparison sites. Separate analysis of important subgroups, such as boys versus girls, or certain diagnoses, such as conduct disorder or depression, will be foregone because too many subjects would be needed to have any assurance of having interpretable results given the predicted effect size.
5. A powerful repeated-measures analysis of variance will be conducted, improving precision by using each subject as his/her own control, to see whether the Demonstration group improves more over time.
6. Since the quasi-experimental design applied in this Evaluation is unique, standard power curve tables could not be used. Instead, statistical modeling was used. In this model, over 1240 hypothetical complete data sets were computer-generated according to the statistical assumptions stated above. Each "model" data set was analyzed with a repeated measures variance analysis.

Results of Power Calculations

The power calculations produced the following results:

<u>Plan</u>	<u>Number of cases after attrition</u>	<u>Statistical Power</u>
Short-term plan	299 Demonstration 150 Comparison	50%
Longer plan	426 Demonstration 361 Comparison	80%

This result means that there is a 50% chance, under the short-term plan, that the statistical analysis will say that the results of the study are due to chance, even if more children improve at the Demonstration site. Hence, with the sample included under the short-term plan, the analyses will be too insensitive to detect true results.

Recommendation

While the short-term plan saves some money, it creates great risk (50%) that an important clinical improvement will be inseparable from the random effects of chance. The longer plan will provide the generally accepted assurance (80%) that the research will have enough data to detect results should they occur. It should be noted, however, that even with this longer plan, this 80% assurance that the effects of the Demonstration can be detected leaves a 20% chance that important effects will be overlooked.

Lampman, C., Durlak, J., & Wells, A. (1992). Statistical Power in Child Psychotherapy Outcome Research. Paper presented at the 1992 American Psychology Association Convention.

Lipsey, M. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage.

APPENDIX B

STATISTICAL POWER IN CHILD
PSYCHOTHERAPY OUTCOME RESEARCH

Claudia Lampman, Joseph Durlak, & Arlene Wells

Paper presented at the 1992
American Psychological Association
Convention

Statistical Power in Child Psychotherapy Outcome Research

Claudia Lampman
University of Alaska, Anchorage
Joseph Durlak and Anne Wells
Loyola University of Chicago



Abstract

A meta-analysis of 377 child psychotherapy outcome studies indicated that effect sizes differed as a function of the type of outcome measure and general form of treatment used. Based on these data, the number of subjects necessary to attain 80% power using various outcome measures and treatments was calculated. The sample sizes needed to achieve adequate power were from two to six times greater than the actual number of subjects typically used in previous child therapy studies. These data underscore the need for researchers to attend to power considerations when planning child therapy outcome studies.

Introduction

Statistical Power in Child Psychotherapy Outcome Research

Statistical power is defined as "... the probability that statistical significance will be attained given that there really is a treatment effect" (Lipsey, 1990, p. 20). In other words, power is the probability of correctly rejecting a false null hypothesis. The likelihood of detecting a treatment effect is associated with many features of a study's design, including the group assignment procedure, the reliability of measures, the fidelity with which treatment is implemented and characteristics of the samples and settings used.

However, even if an experiment is demonstrated to have adequate internal, construct and external validity, it may still fail to be sensitive enough (statistically speaking) to detect a treatment outcome. Statistical power is related to the statistical conclusion validity of a study (Cook and Campbell, 1977), and the critical factor here is sample size.

A number of power surveys have been conducted of various psychological literatures including Cohen's seminal (1962) paper on statistical power in abnormal and social psychological research. In general, these studies have demonstrated that psychological researchers often design, conduct and publish data based on studies with inadequate power (Chase and Chase, 1976; Cohen, 1962; Holmes, 1979; Rossi, 1990). Most of these reviewers have admonished researchers to address the issue of statistical power in the planning phases of research rather than as a post hoc explanation for findings failing to support a treatment's effectiveness. Despite the usefulness of the power reviews, it appears the statistical conclusion validity of studies published in even the most prestigious journals has not improved over the past several decades (Sedlmeier and Gigerenzer, 1989).

One reason that power calculations are not routine procedures in the design of studies may be that an estimate of the expected treatment effect is needed to compute power, along with sample size, designated probability level and directionality of the test (one vs. two-tailed). The power surveys described above typically use a range of sample size to determine the probability that it could detect small, medium and large treatment effects (see Cohen, 1962). Another problem is that the true population effect sizes are unknown, and thus it is difficult to design a study with an adequate, but not excessive sample size. The size of a sample needed to achieve reasonable, say 80 percent, power level differs enormously between small (e.g., .10) and large (e.g., .70) effects. A potential solution to this problem is meta-analysis — which has become an increasingly popular technique for summarizing the findings from a research literature — by providing a more accurate estimate of the true

population effect size than traditional non-quantitative reviews. The results of a meta-analysis can be very useful to researchers planning studies.

The purpose of this paper is to make use of an extensive meta-analytic review of the child psychotherapy literature to present researchers with useful information for the design of child therapy studies.

Method

Meta-analytic Procedures

A total of 367 child psychotherapy outcome studies were reviewed. 287 were journal articles, 14 were book chapters and 66 were unpublished dissertations. Studies eligible for review consisted of reports appearing through the end of 1983 in which some form of psychotherapy for maladapted children (age ≤ 13) was compared with a control group.

Separate effect sizes (ESs) were calculated for each of six categories of outcome measures: behavioral observations, peer sociometrics, measures of academic achievement (standardized test scores or school grades), nonacademic performance measures (e.g., measures of interpersonal problem solving skills and cognitive tempo) and both normed and non-normed rating scales and checklists. Initially 1237 ESs were calculated, however, effects within the same outcome category and same type of treatment were averaged within each study, resulting in a total of 656 ESs that were used in analyses.

Power Analyses

The formula for computing sample size given effect size, alpha and desired power is:

$$n = \frac{n_{.10}}{100 d^2} + 1$$

where $n_{.10}$ is the sample size for the given probability level and desired power when ES is .10 and d = effect size (Cohen, 1977, p. 63).

Results and Discussion

Using the procedures described by Hedges and Olkin (1985), the effect sizes were partitioned into twelve homogeneous clusters based on the type of outcome measure and general form of behavioral treatment used.

Effect sizes for these studies are presented in Table 1, a and with average N per group of the studies in each cell. Table 1, b displays the sample size necessary to achieve 80% power at alpha = .05, one-tailed, given the stipulated treatment effect size and N .

The results of the meta-analysis indicate that the type of measure used to assess outcome and the general type of treatment conducted interact to moderate the impact of child psychotherapy. In fact, twelve distinct clusters of studies were found, with widely varying effect sizes, suggesting that the interpretation of an overall effect size for child psychotherapy would be misleading. The associated sample sizes and power calculations suggest that it would be prudent for researchers planning child psychotherapy outcome studies to think carefully about the selection of outcome measures, as they appear to differ in the ability to detect the effectiveness of various general types of treatment. For example, if an investigator is assessing the effects of nonbehavioral treatment using peer sociometric outcome measures, a sample size of 199 subjects per group is needed to attain 80% power. This estimate is based on a one-tailed test, $\alpha=.05$ to test the difference between

a treatment and control group. This estimate is quite liberal for several reasons. First, it assumes that the treatment group would outperform the control group; a two-tailed test would require even more subjects per group. Second, treatment versus treatment comparisons have been found to yield significantly smaller effects than treatment versus control comparisons (Kazdin & Bass, 1989). Finally, the sample sizes necessary to achieve 80% power also increase as alpha decreases.

The large discrepancies between sample sizes actually used and those necessary to attain an acceptable level of power in the studies shown in Table 1 make it difficult to assess how closely the obtained treatment effect sizes represent true population effects. This, in turn underscores the need for researchers to attend to power considerations when planning therapy outcome studies.

Table 1

Mean effect sizes*, mean sample sizes, and sample sizes necessary to achieve acceptable power ** for twelve homogeneous*** subgroups of child psychotherapy studies.

Type of Outcome Measure	Behavioral Treatment	Nonbehavioral Treatment
Behavioral Observation	Mean ES=.65 (8) Mean N per group=16.7 N for 80% power=30	Mean ES=.25 (34) Mean N per group=55.5 N for 80% power=189
Peer Sociometrics	Mean ES=.43 (16) Mean N per group=16.9 N for 80% power=68	Mean ES=.25 (28) Mean N per group=28.0 N for 80% power=189
Normed Rating Scales	Mean ES=.47 (45) Mean N per group=13.9 N for 80% power=57	Mean ES=.24 (61) Mean N per group=33.0 N for 80% power=216
Non-normed Rating Scales	Mean ES=.62 (61) Mean N per group=17.0 N for 80% power=33	Mean ES=.19 (84) Mean N per group=54.3 N for 80% power=344
Achievement Measures	Mean ES=.45 (38) Mean N per group=24.5 N for 80% power=62	Mean ES=.18 (28) Mean N per group=71.2 N for 80% power=383
Performance Measures	Mean ES=.54 (66) Mean N per group=17.3 N for 80% power=43	Mean ES=.43 (21) Mean N per group=29.1 N for 80% power=68

* all effect sizes differed significantly from zero ($p < .01$); n of studies in parentheses
 ** $\alpha = .05$, one-tailed test, n is per group
 *** each subgroup achieved within group homogeneity of effect sizes ($p < .01$)

References

- Chase, L.J. & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, 61 (2), 234-237.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65 (3), 145-153.
- Cohen, J. (1977). *Statistical Power for the Behavioral Sciences*. New York: Academic Press.
- Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues of Field Settings*. Chicago: Rand McNally.
- Holmes, C.B. (1979). Sample size in psychological research. *Perceptual and Motor Skills*, 49, 283-288.
- Kazdin, A.E. & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, 57, 138-147.
- Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage.
- Rossi, J.S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58 (5), 646-656.
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, 105 (2), 309-316.

APPENDIX C

POWER ANALYSIS DISCUSSION AND DOCUMENTATION

FROM MATERIAL SUBMITTED BY
VANDERBILT UNIVERSITY
APRIL 30, 1993

**FT. BRAGG EVALUATION
ANALYSIS PLAN**

DRAFT

APRIL 30, 1993

Power Analysis

Power analysis is done when planning a study in order to determine how many subjects are needed for adequate statistical power. Power is the ability to detect differences when they actually occur. A power analysis is not normally done by analyzing real data because it is not legitimate to look at data and then stop gathering when the desired results occur. This is true because standard statistical tests all assume that the test is done once then reported; to use non-standard procedures with these tests would seriously hurt their accuracy.

The Monte Carlo power analysis was based on a simplification of the DMM analysis described above, viz. univariate repeated measures analysis. This simpler analysis is more powerful (requires fewer subjects) than the doubly multivariate analysis because fewer variables are used and also because fewer parameters had to be estimated by the simpler analysis. (The ANOVA assumes correlations are uniform; the MANOVA estimates them.)

In comparing the power of this simple repeated measures design to univariate repeated MANOVA, we found that MANOVA costs about a 5% loss of power (or roughly 100 more cases). Thus the single variable repeated measures ANOVA is a conservative test, telling us we need fewer subjects than we actually need for adequate statistical power (80% chance of finding an effect given an effect exists).

Using computer generated data surely would sound strange to the nonstatistician, but "Monte Carlo" simulations are the standard method used by statisticians to test statistical ideas when the problem is too complicated to describe by purely theoretical equations. In the Monte Carlo power analysis, computer generated data was examined to make sure the mean, standard deviation, and the cross-wave correlations were correct. Then the "data" were analyzed in repeated measures ANOVA or univariate MANOVA. By repeating this process hundreds of times and then keeping score on the results we could see what actually happened when we analyze data like the Ft. Bragg Demonstration's. Trying the analysis with varying numbers of "subjects" permits us find out how many subjects were needed for 80% power. If we peeked prematurely at the real data in order to decide when we had enough subjects, we would have ruined the chance to use standard statistical estimates in the way that they were designed.

**FT. BRAGG EVALUATION
DOCUMENTATION FOR POWER ANALYSIS**

APRIL 30, 1993

Basic data

12:53 Wednesday, April 7, 1993

	Demo vs. Comparison					
	Comp			Demo		
	MEAN	STD	N	MEAN	STD	N
Wave 1 CBCL score	66.12	10.00	30000	66.05	10.00	50000
Wave 2 CBCL score	64.48	10.01	30000	63.19	10.00	50000
Wave 3 CBCL score	62.95	9.37	30000	60.52	9.46	50000

The CBCL has a mean of 50 SD 10 for normal children. Ours are in the mid 60's. The power analysis assumes that

Basic data

12:53 Wednesday, April 7, 1993

CORRELATION ANALYSIS

3 'VAR' Variables: CBCLXIJ1 CBCLXIJ2 CBCLXIJ3

Simple Statistics

Variable	N	Mean	Std Dev	Sum
CBCLXIJ1	80000	66.0775	9.9994	5286198
CBCLXIJ2	80000	63.6737	10.0209	5093899
CBCLXIJ3	80000	61.4283	9.4999	4914260

66.0 66
64.5 63
63.0 60

Simple Statistics

Variable	Minimum	Maximum	Label
CBCLXIJ1	23.4102	113.2344	Wave 1 CBCL score
CBCLXIJ2	21.0078	104.3438	Wave 2 CBCL score
CBCLXIJ3	21.6016	100.9531	Wave 3 CBCL score

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 80000

	CBCLXIJ1	CBCLXIJ2	CBCLXIJ3
CBCLXIJ1	1.00000	0.49971	0.26472
Wave 1 CBCL score	0.E+00	0.E+00	0.E+00
CBCLXIJ2	0.49971	1.00000	0.42365
Wave 2 CBCL score	0.E+00	0.E+00	0.E+00
CBCLXIJ3	0.26472	0.42365	1.00000
Wave 3 CBCL score	0.E+00	0.E+00	0.E+00

12:53 Wednesday, April 7, 1993

We assume time and treatment makes everyone (average) .3 SD better. The demo provides an additional 0.25 by fitting Tx better to more children.

UNIVARIATE PROCEDURE

Variable=CBCLXIJ1 Wave 1 CBCL score

Moments

N	80000	Sum Wgts	80000
Mean	66.07747	Sum	5286198
Std Dev	9.999381	Variance	99.98762
Skewness	-0.00057	Kurtosis	-0.0047
USS	3.573E8	CSS	7998910
CV	15.13281	Std Mean	0.035353
T:Mean=0	1869.069	Prob> T	0.E+00
Sgn Rank	1.6E9	Prob> S	0.E+00
Num = 0	80000		

Quantiles(Def=5)

100% Max	113.2344	99%	89.39063
75% Q3	72.8125	95%	82.39844
50% Med	66.125	90%	78.82813
25% Q1	59.3125	10%	53.20313
0% Min	23.41016	5%	49.53125
		1%	43.02344
Range	89.82422		
Q3-Q1	13.5		
Mode	66.76563		

Extremes

Lowest	Obs	Highest	Obs
23.41016(74781)	103.6875(43762)
25.33594(55893)	105.1875(33112)
26.26953(70713)	105.6563(79624)
28.03516(54923)	106.875(33460)

28.66406(27988) 113.2344(28944)

The between-wave cross correlations are about $r=0.50$ for adjacent waves, about $r=0.25$ for nonadjacent waves.

Basic data

12:53 Wednesday, April 7, 1993

UNIVARIATE PROCEDURE

Variable=CBCLXIJ2 Wave 2 CBCL score

Moments

N	80000	Sum Wgts	80000
Mean	63.67374	Sum	5093899
Std Dev	10.02092	Variance	100.4189
Skewness	-0.00888	Kurtosis	0.000346
USS	3.3238E8	CSS	8033413
CV	15.73792	Std Mean	0.035429
T:Mean=0	1797.205	Prob> T	0.E+00
Sgn Rank	1.6E9	Prob> S	0.E+00
Num != 0	80000		

Quantiles(Def=5)

100% Max	104.3438	99%	87.10938
75% Q3	70.4375	95%	80.17188
50% Med	63.72656	90%	76.51563
25% Q1	56.94531	10%	50.80469
0% Min	21.00781	5%	47.1875
		1%	40.21875
Range	83.33594		
Q3-Q1	13.49219		
Mode	66.17188		

Extremes

Lowest	Obs	Highest	Obs
21.00781(27417)	102.2188(24285)
21.67969(41013)	102.2188(61906)
23.29688(61119)	102.6563(9283)
23.60547(54923)	103.1719(78443)
23.98828(39869)	104.3438(14973)

Basic data

12:53 Wednesday, April 7, 1993

UNIVARIATE PROCEDURE

Variable=CBCLXIJ3 Wave 3 CBCL score

Moments

N	80000	Sum Wgts	80000
Mean	61.42825	Sum	4914260
Std Dev	9.499866	Variance	90.24745
Skewness	-0.01171	Kurtosis	-0.01007
USS	3.0909E8	CSS	7219706
CV	15.46498	Std Mean	0.033587
T:Mean=0	1828.924	Prob> T	0.E+00
Sgn Rank	1.6E9	Prob> S	0.E+00
Num != 0	80000		

Quantiles(Def=5)

100% Max	100.9531	99%	83.52344
75% Q3	67.89063	95%	76.98438
50% Med	61.42188	90%	73.60938
25% Q1	55	10%	49.26563
0% Min	21.60156	5%	45.78516
		1%	39.20703
Range	79.35156		
Q3-Q1	12.89062		
Mode	64.20313		

Extremes

Lowest	Obs	Highest	Obs
21.60156(16801)	96.45313(78827)
23.44531(4018)	97.20313(77062)
25.29297(42150)	98.20313(2814)
25.29688(35497)	99.15625(44423)
26.17188(63423)	100.9531(2098)

Basic data

12:53 Wednesday, April 7, 1993

General Linear Models Procedure
Class Level Information

Class Levels Values

SITE 2 Comp Demo

Number of observations in data set = 80000

Basic data

12:53 Wednesday, April 7, 1993

General Linear Models Procedure
Repeated Measures Analysis of Variance
Repeated Measures Level Information

Dependent Variable CBCLXIJ1 CBCLXIJ2 CBCLXIJ3

Level of WAVE 1 2 3

Basic data

12:53 Wednesday, April 7, 1993

General Linear Models Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SITE	1	90132	90132	521.62	4E-115
Error	79998	13823067	173		

Basic data

12:53 Wednesday, April 7, 1993

General Linear Models Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source: WAVE

DF	Type III SS	Mean Square	F Value	Pr > F	Adj G - G	Pr > F	H - F
2	71129.5857	355764.7928	6129.16	0.E+00	0.E+00	0.E+00	

Source: WAVE*SITE

DF	Type III SS	Mean Square	F Value	Pr > F	Adj G - G	Pr > F	H - F
2	51927.2900	25963.6450	447.	2E-194	4E-187	4E-187	

Source: Error(WAVE)

DF	Type III SS	Mean Square
159996	9286901.3432	58.0446

Greenhouse-Geisser Epsilon = 0.9614
Huynh-Feldt Epsilon = 0.9615

NOTE: Copyright(c) 1985,86,87 SAS Institute Inc., Cary, NC 27512-8000, U.S.A.
NOTE: SAS (r) Proprietary Software Release 6.04
Licensed to VANDERBILT UNIVERSITY, Site 11765001.

NOTE: AUTOEXEC processing completed.

```

1
2
3
4 * Set numbers prior to running a problem
5 * *****
6 %let NDemo = 50000;%let NComp = 30000;
7 %let VarDemo = -0.25 ;
8 %let varTime = -0.30;
9 data ALL lies(keep = SITE CBCLxij1 CBCLxij2 CBCLxij3);
10 * *****constants;
11 sqrt2 = sqrt(2.0);
12 * *****make scores;
13 /*
14 varwithin variance within subjects
15 varandm1 random err. on measurement 1
16 varandm2 random err. on measurement 2
17 varandm3 random err. on measurement 3
18 rannor(seed) sas random function mean = 0, sd = 1
19 CBCLxij3 = CBCL score, Xsub ij on third occasion
20 = half of (variance within + random3) + effect of time
21 + effect of demonstration
22 */
23
24 do i = 1 to &NDemo by 1;
25 SITE = "Demo";
26 varwithin = rannor(0);
27 varandm1 = rannor(0);
28 varandm2 = rannor(0);
29 varandm3 = rannor(0);
30
31 CBCLxij1 = (varwithin + varandm1)/sqrt2;
32 CBCLxij2 = (varwithin + varandm2)/sqrt2 + &VarTime/2 +
&VarDemo/2;
33 CBCLxij3 = (0.5*varwithin + 0.3*varandm2 + 1.20*varandm3)/sqrt2
34 + &VarTime + &VarDemo;
35
36 CBCLxij1 = (10.0 * CBCLxij1) + 66.0 ; /* mean 66, Sd 10 */
37 CBCLxij2 = (10.0 * CBCLxij2) + 66.0 ;
38 CBCLxij3 = (10.0 * CBCLxij3) + 66.0 ;
39 output;
40 end;
41
42 do i = 1 to &NComp by 1;
43 SITE = "Comp";

```

If wave*site p<0.05
Then Tx had an effect.

← Make cases of any N desired

} Use random numbers

} To generate a stat.
model to produce
desired means, variance,
cross correlations.

DEMO (c-6) d

```

44      varwithn = rannor(0);
45      varandm1 = rannor(0);
46      varandm2 = rannor(0);
47      varandm3 = rannor(0);
48
49
50      CBCLxij1 = (varwithn + varandm1)/sqrt2;
51      CBCLxij2 = (varwithn + varandm2)/sqrt2 + &VarTime/2 ;
52      CBCLxij3 = (0.5*varwithn + 0.3*varandm2 + 1.20*varandm3)/sqrt2
53                + &VarTime ;
54
55      CBCLxij1 = (10.0 * CBCLxij1) + 66.0 ;      /* mean 66, Sd 10 */
56      CBCLxij2 = (10.0 * CBCLxij2) + 66.0 ;
57      CBCLxij3 = (10.0 * CBCLxij3) + 66.0 ;
58      output;
59      end;
60
61      attrib SITE          label = 'Demo vs. Comparison'
62      CBCLxij1 format = 5.1 length = 3 label = 'Wave 1 CBCL score'
63      CBCLxij2 format = 5.1 length = 3 label = 'Wave 2 CBCL score'
64      CBCLxij3 format = 5.1 length = 3 label = 'Wave 3 CBCL score';
65      run;
NOTE: The data set WORK.ALL_LIES has 80000 observations and 4 variables.
NOTE: The DATA statement used 2.45 minutes.
66
67      options linesize = 72;
68      proc tabulate f = 6.2 data = ALL_LIES;
69      class SITE;
70      var CBCLxij1 CBCLxij2 CBCLxij3;
71      table (CBCLxij1 CBCLxij2 CBCLxij3),
72            (SITE)*(mean*f=6.2 std*f=6.2 N*f=6.0);
73      title 'Basic data';
74      run;
NOTE: The PROCEDURE TABULATE used 1.02 minutes.
75      proc corr;
76      var cbclxij1 -- cbclxij3;
77
78      proc univariate;
NOTE: The PROCEDURE CORR used 34.00 seconds.
79      var cbclxij1 -- cbclxij3;
80
81      options linesize = 72;
82      proc glm;
NOTE: The PROCEDURE UNIVARIATE used 2.15 minutes.
83      classes SITE;
84      model CBCLxij1 CBCLxij2 CBCLxij3 = SITE/nouni;
85      repeated wave 3/nom;
86      run; quit;
NOTE: The PROCEDURE GLM used 2.57 minutes.
87
88      endsas;
NOTE: SAS Institute Inc., SAS Circle, PO Box 8000, Cary, NC 27512-8000

```

No DEMO effect.

Check generated scores for fidelity to assumptions.

Are cross i's right?

Do repeated measures ANOVA with G-G corrections for nonconstancy of covariance.

MANOVA screen would be appropriate but would ↓ power.

EXTERNAL PEER REVIEW
OF THE REPORT CR 93-002

APPENDIX D

The University of Texas
Health Science Center at Houston



SCHOOL OF PUBLIC HEALTH
Health Services Organization

1200 Herman Pressler
P.O. Box 20186
Houston, Texas 77225
(713) 792-4372
(713) 792-4471

May 10, 1993

Edward D. Martin, M.D.
Assistant Secretary of Defense (Health Affairs)
The Pentagon, Washington DC

Dear Dr. Martin:

I have now completed my review of the materials submitted to me on April 23, 1993, by Dr. Scott Optenberg.

In the absence of information on several key factors relevant to the successful execution of a project of this magnitude, it is indeed impossible to conduct an objective evaluation of all the claims of the investigators for the Fort Bragg Demonstration project. I will therefore limit my comments to the power analysis performed by the Army Statisticians in an in house effort to determine if the demonstration project should be extended.

The investigators at Fort Bragg are interested in detecting a standardized difference of .25 between the experimental and control subjects for the short term plan. They anticipate 299 demonstration and 150 control cases at wave 3. As demonstrated by the detailed power analysis developed for this purpose by Dr. Optenberg's group, no matter what assumptions are made on the variances of the two populations, the minimum power that may be attained at wave 3 of the analysis is about 81%. The derivation of the power analysis is based on the theoretical developments presented in Cohen's (1988) book which is regarded as the basic text on power analysis in behavioral sciences.

Similarly, using the anticipated number of cases at the end of (wave 3) the long term plan, (i.e. 426 demonstration and 361 comparison cases) a power of at least 90% will be obtained.

In the Fort Bragg demonstration project, a power of .80 for detection of a relatively small difference (i.e. .25 SD) in improvement between subjects in the experimental and control groups is very impressive considering that most research studies in social sciences are under powered (power <.80) for detecting anything but large differences (Lipsey 1990). Thus, the short-term plan is more than sufficient to meet the objectives of this demonstration project.


The investigators justification for a long term plan is based on the argument that if only the short term plan were to be carried out, the likelihood of detecting a statistical significance in the presence of a treatment effect would be 50%. This claim has not been demonstrated mathematically by the investigators and as shown by the power analysis performed by the army statisticians using appropriate statistical procedures is in serious error.

On reviewing the documentation dated April 30, 1993 from Vanderbilt University (received by me on 5/8/93), some inconsistency in the claims of the Investigators/Evaluators of the demonstration project is apparent. On page 4 of the above document, they state that the project has been losing about 15% of the subjects per wave. Using this attrition rate the 1065 wave 1 cases (demonstration plus control) should result in 1065 (.85)(.85) or 769 cases. Yet under data collection assumptions the 1065 wave 1 cases will result in only 449 (299 demonstration and 150 control) cases. Therefore, the statistical power under the proposed short-term plan may be even higher than 81%.

Furthermore, investigators have repeatedly mentioned not wanting to "peek prematurely" at the real data for fear of "ruining the chance to use standard statistical estimates in the way that they were designed". To obtain the power associated with a study on treatment effectiveness, all one needs is some assumption on the variance of the two treatment outcomes (in this study demonstration and control cases), the number of individuals in each group, the effect size and the level of significance. Power calculation does not require a "peek" at the actual data. Hence the use of Monte Carlo simulation to estimate the power of the study is unnecessary and irrelevant.

If I may be of further help, please feel free to call me at (713) 792-4472.

Sincerely,


Asha S. Kapadia
Professor and Convener
of Biometry

ASK:rf